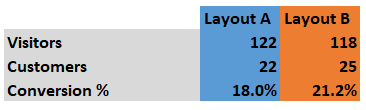
**Introduction**

This post outlines A/B testing, and the steps necessary to plan and build our own A/B test as it really is the industry standard tool for product development. To demonstrate, imagine that we are trying to work out whether rearranging our website into a new format affects conversion rate.

One approach would be to run both versions to selected customers and make a judgement based on these numbers alone:



In this case, we would conclude that layout B is superior to Layout A. However, such a simple approach suffers from two possible errors that we are very familiar with (if we want to investigate why simply choosing a layout plan would cause the following errors, please read the last article)

* Type I error: falsely concluding that our intervention was successful. Also known as a false positive result.
* Type II error: falsely concluding that our intervention was not successful. Also known as a false negative result.

An A/B test will enable us to accurately quantify our effect size and errors, and so calculate the probability that we have made a type I or type II error. Please note that it is very important to understand the true effect size of our test, then we can proceed to making business decisions.

**Principles of A/B tests**

This part of the article will outline the design principles of A/B tests and how to ensure that a trial is effective and cost-efficient. To achieve that, we will be relying on the concepts of statistical significance and statistical power.

We might start thinking about an A/B test based on a question or an idea. For example, would SMS reminders for loan repayments will reduce loan defaults? With some work, we can turn it into a hypothesis and then an A/B test that will evaluate the exact gain that results from the new SMS system.

To do this, we first need to form our question as a hypothesis, we then need to work out our randomization strategy, sample size and finally our method of measurement.

**The hypothesis**

A hypothesis is a formal statement describing the relationship we want to test. A hypothesis must be a simple, clear and testable statement that contrasts a control group (layout A) with a treatment group (layout B)

To form a hypothesis, we re-phrase "does an SMS system improve repayment?" Into two statement, a null hypothesis and an alternative hypothesis:

* Null hypothesis: the null hypothesis usually states that there is no difference between treatment and control groups.
* Alternative hypothesis: the alternative hypothesis states that there is a difference between treatment and control groups.
* Notably, a hypothesis should include reference to the population under study, the intervention, the comparison group, the outcome and the time.

Now let's use these rules (PICOT) to formally design a hypothesis using our layout example

* Null hypothesis: Amazon.com visitors that receive layout B will not have higher conversion rates compares to visitors that receive layout A.
* Alternative hypothesis: Amazon.com visitors that receive layout B will have higher conversion rates compared to visitors that receive layout A.
  + Population: individuals who have visited the Amazon.com site
  + Intervention: new website layout B
  + Comparison: visitors receiving layout A
  + Outcome: conversion rate
  + Time: end of visit to Amazon.com

Now, let's take a minute and reflect on what's considered as bad hypothesis.

* Null: banks with nicer colors will not affect loan repayment
* Alternative: banks with nicer color will affect loan repayment
  + There is no clear definition of "nicer colors", my opinion might be different than yours. This is an example of poor intervention definition.
  + What banks? Where, and what level? Do we mean bank branches? If so, what's the region that our samples are taken from. This is a poor population specification.
  + How are we measuring this? Loan default rates, days past due, total branch losses? This is an example of poor outcome specification.

A strong hypothesis will hold the A/B test together and provide guidance on the design and analysis.

**Randomization**

Return to the Amazon example, once we have a well formed hypothesis we can think about randomization strategies. To extend our example from above, we could randomize our visitors in two ways:

* Randomly assign visitors to layout A or B
* Allow visitors to opt-in to new layout betas

Let's first examine the reasons why we randomize in an A/B test.

* Distributing co-variance evenly
  + Co-variance are factors that might influence your outcome variable, for example, visitor geolocation, gender and risk-appetite.
* Eliminating statistical bias
  + Statistical bias can occur when our sample is substantially different from our target population. We originally assume that our sample is representative of our population, deviations from this assumption can lead to invalid test results.

However, also note that there are draw backs for each randomization method as well.

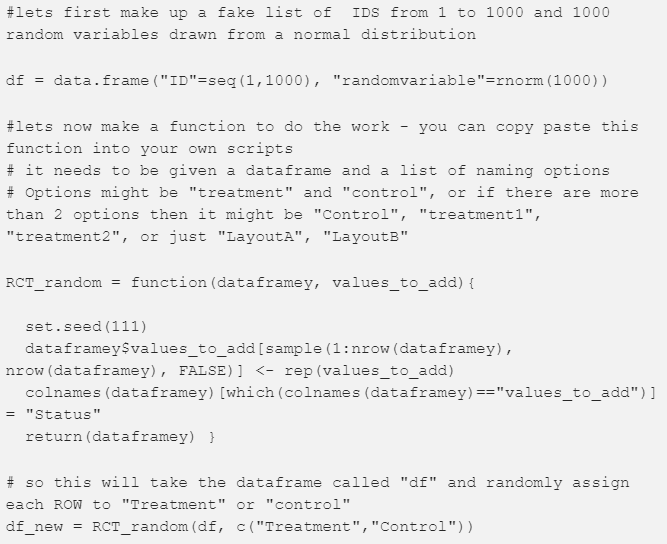
* Randomization bias: bias due to poor randomization resulting in unbalanced Treatment/Control groups. This allows some co-variance to exert more influence in one group than another.
* Selection bias: bias would also result if we were to allow visitors to assign themselves to Treatment/Control groups. This is because there may be unobservable co-variates that are associated with a Treatment/Control choice. For example, visitors with more risk appetite might select themselves into the beta testing group.

Both of these will lead to what is called confounding bias, this means it will be difficult to untangle effects that are due to poor randomization vs. Effects that are due to the actual intervention.

* If each participant has an equal chance of being randomly assigned to a treatment/control group then randomization will be free of bias. This will result in both observable and unobservable co-variates being spread equally to both groups.

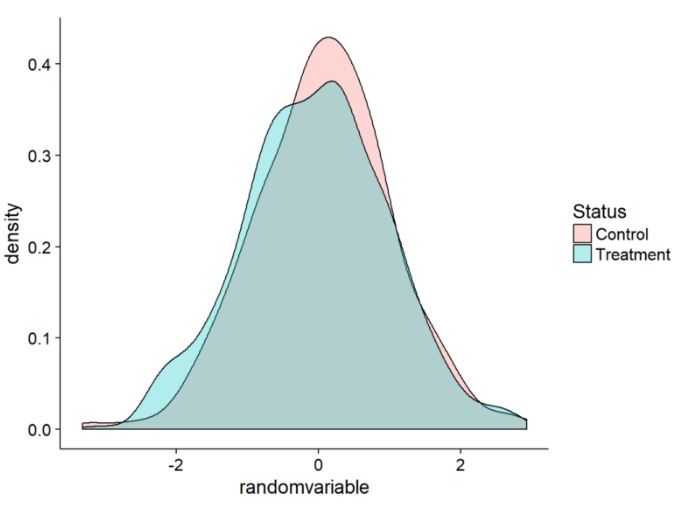
So which of these now seems like a better strategy from a statistical point of view?

* While there might be some organizational reasons why we can't do the former strategy, it is certainly a more statistically robust trial.
* This is because we have complete control. If we used strategy 2, allowing visitors to opt-in to new layouts, then there would likely be unobservable factors at play that confound effects from unobservable factors with our outcome variable.
* Therefore, it is important to select treatment and control group entirely randomly.



Note that we should always double-check our randomization to ensure it has proceeded as expected, we can look at the distributions of the most important key variables. It is also possible to run an appropriate hypothesis test to assess whether the distribution are different.

* For the hypothesis testing, we need to set our p-value to 0.01 due to the multiple comparison problem.



**Cluster Randomization**

Sometime, we won't be able to randomize at the individual level but we will still want to measure at that level.

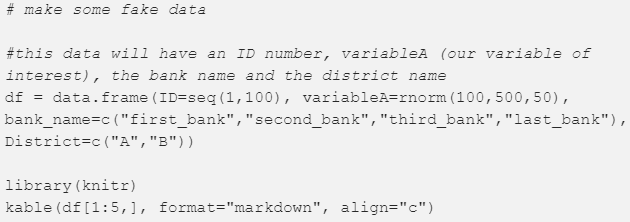
* For example, randomly assess the impact of incentives for bank staff on customer loan repayment rates.
* It not make sense to randomize at the customer level in this example, we would instead need to randomized at the bank level whilst still measuring the customer level outcome.

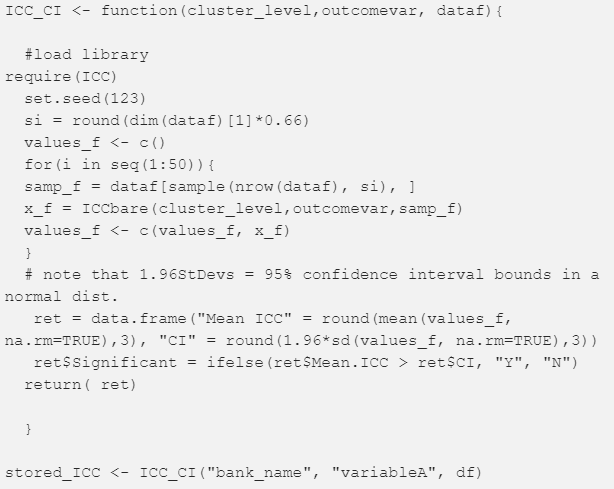
The process of randomizing at one level but measuring at another causes complications in A/B testing design. The solution is to use a cluster randomized trial or cluster A/B test.

A cluster randomized trial is similar to an A/B test but our unit of randomization becomes the cluster rather than the individual (so the bank is the cluster).

* We can measure clustering with the intra-cluster correlation or ICC which will tell us how correlated the response of individuals within a cluster are.
* ICC runs from 0 to 1.
* A higher correlation causes more analysis complication, so an ICC of 0 would be ideal.

Let's first look at how we can calculate ICCs:

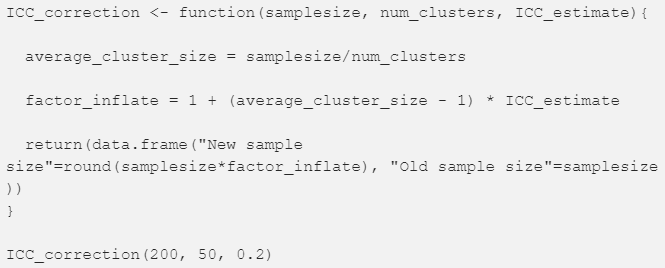




We can see from this calculate that our ICC between customers in the same bank is (0.022 +/- 0.081). This result can be interpreted as the banks account for a negligible amount of similarity between customers.

However, if we have a significant ICC then we will need to adjust our trial design.

* Option 1: calculate a summary metric for each cluster (cluster mean). Each cluster then provides only one data point and allows us to continue with the assumption that our data is independent, we can then proceed with standard statistical tools as normal.
  + So if we have 500 individuals in 45 groups, we end up with 45 data points. This means our power, sample size and analysis calculations also need to be carried out at the cluster level. It means that we can simply ignore ICC from here on out.
* Option 2: use the ICC with the number of individuals to work out our new sample size.



So an initial sample size of 200 customers, with 30 clusters and an ICC of 0.2 would lead to a new sample size of 320. Note how a relatively small ICC increases cause our sample size by more than 50%.

**P-value Threshold, or Alpha**

Once we have a testable hypothesis and a robust randomization strategy, we will need to test the effectiveness of an intervention with a hypothesis test. This hypothesis test will yield a p-value, which is the probability that our data could generated purely by chance. In other words, the probability a false positive result occurring.

* When using a hypothesis test we must set an acceptable rate of false positives, aka an alpha level.
* The most common p-value threshold is 0.05. This means that we are willing to accept a 5% risk of generating a false positive and wrongly concluding that there is a difference between our treatments when in fact, there is not.
* For an intervention that might have adverse effects on customers, we might want to set a threshold of 0.01 to ensure the positive effects to be correct.

**Sample Size Calculation**

Hypothesis framing, randomization process and p-value threshold, are all essential pieces for a robust A/B test. But they are more leaning towards the strategy side of the experiment. Now, we are left with more field relevant problems to think about, and they are minimum detectable effect, effect size, statistical power, sample sizes.

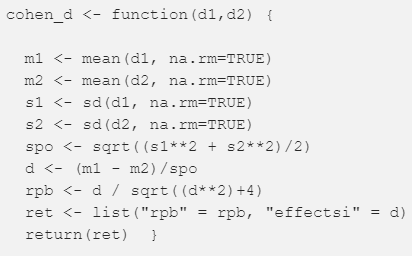
**Before Calculating Sample Sizes**

The main piece of information needed before a sample size calculation is an estimate of intervention effect size. There are generally two ways we can derive an effect size estimate for our calculations:

* Using the Cohen's D equation
* Set a minimum detectable effect
* MDE -> effect size ->  samples size -> statistical power

Effect size is a simple way to measure the effect of a treatment. It is basically the difference in means between samples divided by the standard deviation of the samples. So an effect size of 1 means the data has shifted over an amount equal to 1 standard deviation.

Cohen's D



MDE

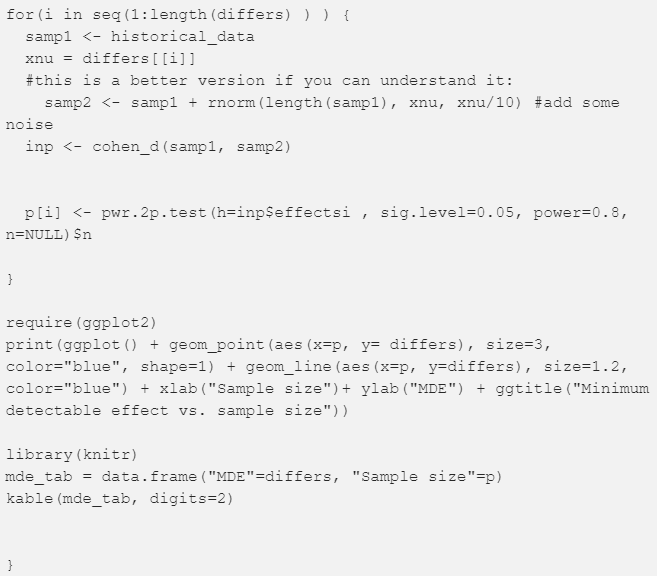
Perhaps the most business focused approach is to use the minimum detectable effects.

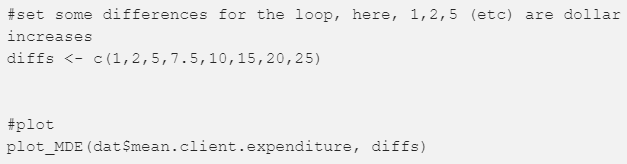
The MDE approach will ask "what is the minimum effect that I would need to see for the intervention to be worthwhile?", we would then set the effect size to that value.

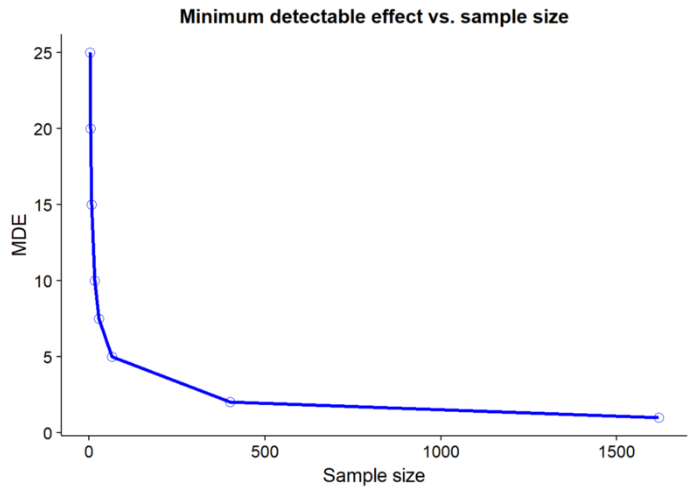
* For example, implementing layout B across Amazon.com might be quite expensive. We could calculate that the implementation would cost about $20,000 of staff time.
* We would only care about being able to detect effects greater than $20,000.
* Let's say that $20,000 corresponds to a 2% bump in conversion rate, we then set our MDE to 2 points and our effect size to 2 points for our calculations.

The MDE approach can be very powerful. It is used to estimate sample size in the pre-analysis and post-analysis to say what size effect we would have been able to detect with a power of 0.8.





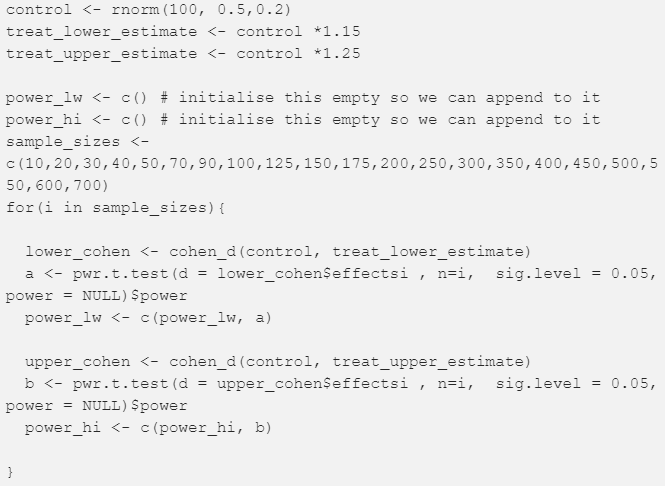


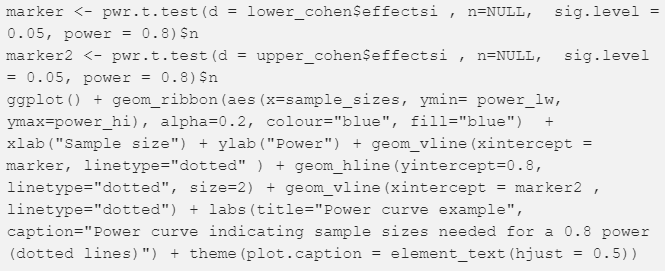


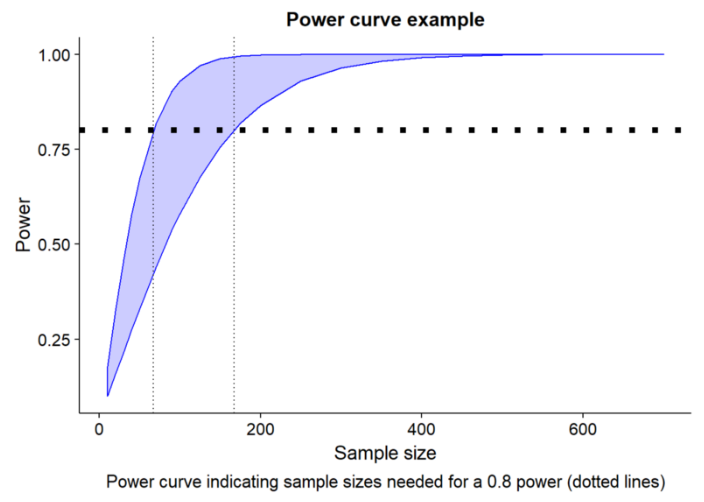
**Calculating Sample Sizes**

Let's think of an example. Imagine we want to access the effect of an new interest rate on fertilizer adoption.

* Our control group has the same interest rate as before, whereas, our test group has an interest rate reduced by 3 points.
* Our understanding of the customer means we think we will see a 15 to 25% increase in fertilizer adoption with the treatment.

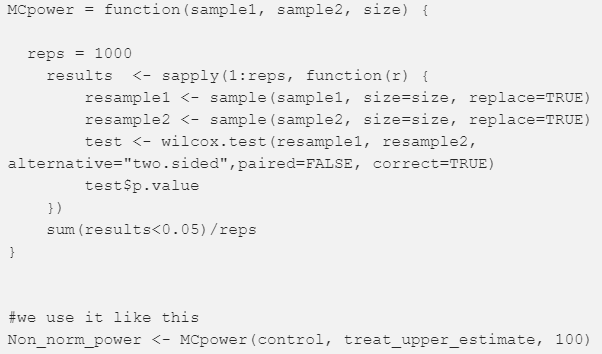






The above plot shows that it would require 400 samples (200 control, 200 test) in this experiment to detect a 15% change in adoption with a power of 0.8. But what about non-normal powers?

* The function below will calculate the power for a non-normal distribution. The function requires data for sample 1 and 2 and a sample size.

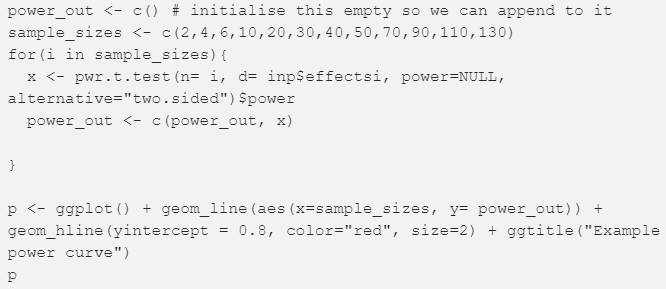


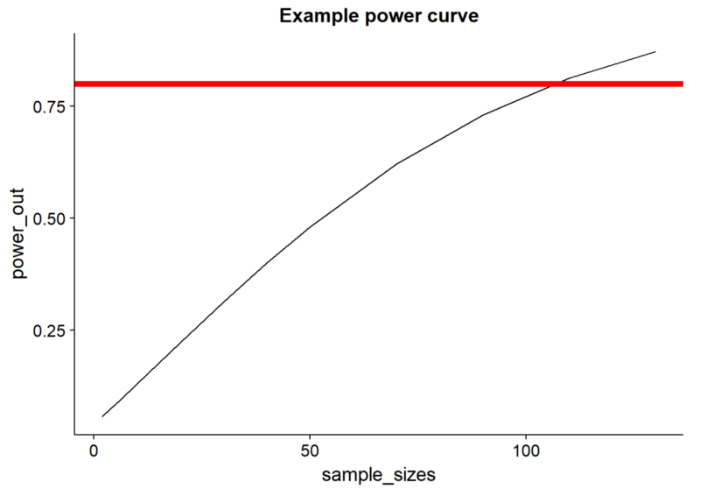
The Best Practices When Analyzing an RCT:

* Even though we calculate the sample size before an RCT, it is important to calculate the power (which is related to sample size) after the RT to make sure our study has sufficient power.
  + Lower power means we are unlikely to detect an effect from an intervention, an thus would make our results less reliable.
* We should always plot density plots or boxplots for each key variables.
* We should run a Shapiro test on our data to make sure it is normal before deciding what method to use
  + T-test: assumes a normal-like distribution
  + Paired T-test: a t-test used for measuring the same subject
  + Two-sample T-test: for two different groups
  + Welch test: an improvement on the t-test, which is more robust to unequal variances.
  + Wilcoxon-test: a good test for non-normal data with very few assumptions.
    - Samples are random
    - Samples are independent
    - Values have an order

**After Calculating Sample Sizes**

We can make a loop to look at the power with different sample sizes, and then plot this with a line indicating where a power of 0.8 is:





**Measurement**

The final part of the A/B test is the measurement itself. This is often a neglected part of test design and therefore a major source of later analysis issues. Any hypothesis needs to have a good measurement strategy to be useful.

It is vital to spend some time thinking about the caveats and weaknesses of measurement strategies, and then to try to mitigate those weaknesses as much as possible.

Finally, once the data is collected and analyzed, it’s important to make it accessible and reproducible. This means writing clean code (ideally in R-markdown or Jupyter notebook) and saving the raw data on company servers.